

Performance Analysis of an Unreliable $M/G/1$ Retrial Queue with Coupled Switching

Muthukrishnan Senthil Kumar^a, Aresh Dadlani^{b,*}, Kiseon Kim^b

^a*Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore 641-004, India*

^b*School of Information and Communications, Department of Nanobio Materials and Electronics, Gwangju Institute of Science and Technology, Gwangju 500-712, Korea*

Abstract

We investigate the stationary characteristics of an $M/G/1$ retrial queue where the server, subject to active failures, primarily attends incoming calls and directs outgoing calls only when idle. On finding the server unavailable (busy or failed), inbound calls join the orbit and reattempt for service at exponentially-distributed time intervals. The system stability condition and probability generating functions of the number of calls in orbit and system are derived and evaluated numerically in the context of mean system size, server availability, failure frequency, and orbit waiting time.

Keywords: Retrial queue, server breakdown, coupled switching, reliability, stationary distribution

1. Introduction

Blended call centers have recently evolved as an effective and profitable communication asset in bridging companies and their customers. Such communication systems are capable of managing a mixture of both, inbound and outbound call operations that require instant service [1]. Outbound calls are made by the server only when there are no inbound calls in the system. This feature, commonly also referred to as *coupled switching* or *two-way communication*, increases productivity by reducing the idle time experienced by the service agents [2]. Besides, incoming calls that find the server busy have an intrinsic tendency to retry for service after some random time [3, 4]. As a result, stochastic behavior analysis of coupled switching in call centers under the influence of retrying customers is crucial to statistical practitioners and network service managers.

There exist a number of seminal works dedicated to coupled switching in the retrial queueing literature. In [5], the authors analyzed some expected perfor-

*Corresponding author

Email addresses: msk@amc.psgtech.ac.in (Muthukrishnan Senthil Kumar), dadlani@gist.ac.kr (Aresh Dadlani), kskim@gist.ac.kr (Kiseon Kim)

15 mance measures of the $M/G/1/K$ priority retrial queue with coupled switching
under the assumption that incoming and outgoing calls follow the same service
distribution. Nevertheless, such an assumption limits the practicality of the
model as customers may have different service needs. Although the authors of
[6] derive the first partial moments for a $M/G/1/1$ retrial queue model with
20 different service time distributions using mean value analysis, it cannot be used
to obtain the stationary distribution. In-depth analysis of the $M/M/1/1$ retrial
queue with coupled switching and different service time distributions for single
and multiple server cases have been reported in [7]. Artalejo *et al.* [8] proposed
an embedded Markov chain approach to study the steady-state behavior of a
25 couple-switched $M/G/1$ retrial queue with tailed asymptotic analysis of number
of customers in the orbit. Nonetheless, in practice, the server may experience
multiple failures which, to our best knowledge, has not been scrutinized in mod-
eling systems with two-way communication. Hence, continuous-time analytical
characterization of an unreliable single server retrial queue with coupled switch-
30 ing under steady state is imperative from the viewpoint of both, queueing as
well as reliability analysis.

The immediate goal of this letter is to study the impact of server failure on
the performance of the $M/G/1$ queue with two-way communication having an
infinite orbit and generally distributed server repair time. In particular, we ob-
35 tain the system stability condition using the embedded Markov chain technique,
followed by the supplementary variable approach to obtain in closed-form the
probability generating functions (pgfs) for incoming calls in orbit and the sys-
tem. Furthermore, we conduct numerical simulations for various performance
metrics to corroborate our theoretical analysis.

40 2. System Model Formulation

We consider a single server retrial queue in which the primary inbound calls
follow a Poisson arrival process with rate λ . If the server is idle, it makes an
outgoing call in exponentially distributed time with rate α . As in reality, the
time taken to serve incoming and outgoing calls is assumed to be different. If an
45 incoming call finds the server busy, it then enters the orbit and re-attempts to
seek service after an exponentially distributed time with rate ν . Otherwise, the
incoming call commences service immediately. Since the server may breakdown
while serving calls, without loss of generality, we assume that the lifetime of
the server follows an exponential distribution with rates β_1 and β_2 during the
50 service of inbound and outbound calls, respectively. On failure, the server is
instantly sent for repair which has a generally distributed time.

For the sake of consistency, we define $i \in \{1, 2\}$ to differentiate between
incoming and outgoing calls. Henceforth, $i = 1$ refers to incoming calls, while $i =$
2 indicates outgoing calls. Let $S_i(x)$ and $R_i(x)$ be the cumulative distributions of
service and repair times of i -type calls, respectively. Similarly, let $s_i(x)$ and $r_i(x)$
denote respectively, the probability density functions of service and repair times
of i -type calls. The Laplace transform (LT) of the service and repair times for
each type of call is denoted as $\tilde{S}_i(\theta)$ and $\tilde{R}_i(\theta)$, respectively. We also define $S_i^o(x)$

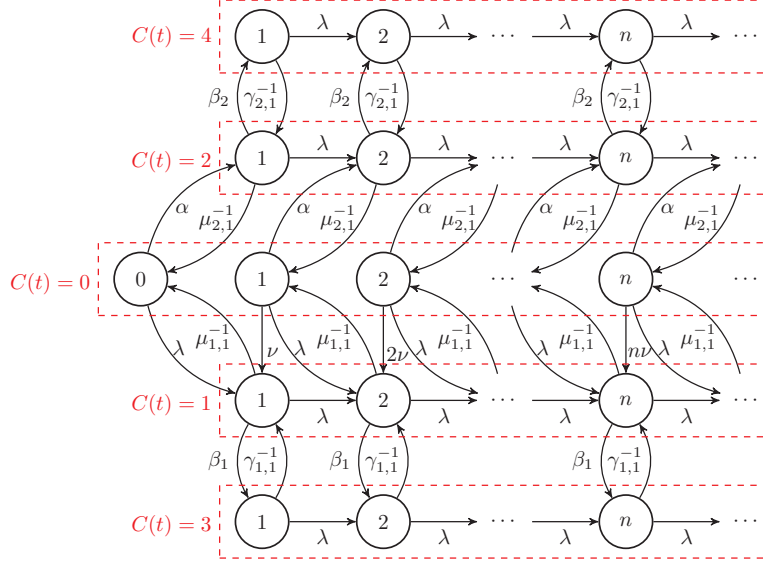


Figure 1: State transitions of the proposed system model.

and $R_i^o(x)$ as the remaining service and repair times, respectively. Moreover, let $\mu_{i,k}$ and $\gamma_{i,k}$ denote the k^{th} moment of service and repair times, respectively. Additionally, the arrival flows of incoming calls, outgoing calls, service time, repair time, and intervals between successive re-attempts are all assumed to be mutually independent. Finally, let $N(t)$ be the number of incoming customer calls in orbit at time t , $M(t)$ be the total number of customers in the system at t , and $C(t)$ be the server state defined as follows:

$$C(t) = \begin{cases} 0, & \text{if the server is } \textit{idle} \\ 1, & \text{if the server is } \textit{busy with incoming calls} \\ 2, & \text{if the server is } \textit{busy with outgoing calls} \\ 3, & \text{if the server } \textit{fails while serving incoming calls} \\ 4, & \text{if the server } \textit{fails while serving outgoing calls} \end{cases} \quad (1)$$

Figure 1 shows the state transitions of $\{(C(t), N(t)); t \geq 0\}$ with state space $S = \{0, 1, 2, 3, 4\} \times Z_+$ for exponential service and repair times. Here, n is the number of incoming calls in orbit and each row represents the server state as defined in (1). For generally distributed service and repair times, the system state at time t can be expressed as the following continuous-time Markov chain:

$$K(t) = \{(C(t), N(t), S_1^o(t), S_2^o(t), R_1^o(t), R_2^o(t)); t \geq 0\}. \quad (2)$$

Based on $K(t)$, we now define the state probabilities as follows where $n, x, y \geq 0$:

$$\begin{cases} P_{0,n}(t) = Pr[C(t)=0, N(t)=n], \\ P_{1,n}(x,t)dx = Pr[C(t)=1, N(t)=n, x < S_1^o(t) \leq x+dx], \\ P_{2,n}(x,t)dx = Pr[C(t)=2, N(t)=n, x < S_2^o(t) \leq x+dx], \\ P_{3,n}(x,y,t)dy = Pr[C(t)=3, N(t)=n, S_1^o(t)=x, y < R_1^o(t) \leq y+dy], \\ P_{4,n}(x,y,t)dy = Pr[C(t)=4, N(t)=n, S_2^o(t)=x, y < R_2^o(t) \leq y+dy]. \end{cases} \quad (3)$$

Here, $P_{0,n}(t)$ is the probability that the server is idle and there are n calls in the orbit at time t . Similarly, $P_{i,n}(x,t)dx$ denotes the joint probability that the server is busy with an i -type call during the remaining service time $(x, x+dx)$ and there are n calls in the orbit at epoch t . For $j \in \{3, 4\}$, $P_{j,n}(x,y,t)dy$ refers to the joint probability that at time t there are n calls in the orbit, the remaining service time is x , and the failed server is fixed within the remaining repair time $(y, y+dy)$ while it was serving an inbound ($j=3$) or an outbound ($j=4$) call.

3. Steady-state Distribution

To identify the pgfs of orbit size and number of calls in the system, we first determine the stability condition of the system using the following theorem.

Theorem 1. *The necessary and sufficient condition for system stability is given by the inequality $\lambda\mu_{1,1}(1 + \beta_1\gamma_{1,1}) < 1$.*

Proof. Let \hat{X}_n be the service completion time of the n^{th} call which includes possible down times (due to server failure) while providing service. For the sufficient condition, we need to prove the ergodicity of $\{L_n; n \geq 1\}$, where $\{L_n\}$ is an irreducible and aperiodic discrete-time Markov chain of $K(t)$ in (2) and is defined as $L_n = N(\hat{X}_n^+)$. Using Foster's criterion and undertaking the same approach as in [8], $\{L_n\}$ is positive recurrent if $|\eta_k| < \infty$ and $\lim_{k \rightarrow \infty} \sup\{\eta_k\} < 0$ for all k , where $\eta_k = E[(L_{n+1} - L_n)/L_n = k]$. This definition of η_k results in:

$$\eta_k = \frac{kv[\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1})-1]}{\lambda+kv+\alpha} + \frac{\lambda[\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1})]}{\lambda+kv+\alpha} + \frac{\alpha[\lambda\mu_{2,1}(1+\beta_2\gamma_{2,1})]}{\lambda+kv+\alpha}. \quad (4)$$

Obviously, if $\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1}) < 1$, then for all k , $\eta_k < \infty$ and $\lim_{k \rightarrow \infty} \sup\{\eta_k\} < 0$, which proves the sufficiency. As pointed out in [9], the non-ergodicity of $\{L_n\}$ can be guaranteed if Kaplan's condition is satisfied, i.e. there exists some $k_0 \in \mathbb{Z}_+$ such that $\eta_k \geq 0$ for $k \geq k_0$ and $\eta_k < \infty$ for all $k \geq 0$. In our case, this condition is satisfied as $r_{i,j} = 0$ for $j < i - 1$, where $P = [r_{i,j}]$ is the one step transition probability matrix. Therefore, $\lambda\mu_{1,1}(1 + \beta_1\gamma_{1,1}) > 1$ implies the non-ergodicity of $\{L_n; n \geq 1\}$, which completes the proof. \square

Using supplementary variable technique, we obtain the following balance

and $h_i(z) = \lambda + \beta_i - \lambda z - \beta_i \tilde{R}_i(\lambda - \lambda z)$:

$$\begin{cases} P_0(z) = \frac{1 - \lambda\mu_{1,1}(1 + \beta_1\gamma_{1,1})}{1 + \alpha\mu_{2,1}(1 + \beta_2\gamma_{2,1})} \phi(z), \\ \tilde{P}_1(z, 0) = \frac{[\lambda(1 - z) + \alpha(1 - \delta_2(z))][1 - \delta_1(z)]}{[\delta_1(z) - z]h_1(z)} P_0(z), \\ \tilde{P}_2(z, 0) = \frac{\alpha[1 - \delta_2(z)]}{h_2(z)} P_0(z), \\ \tilde{\tilde{P}}_3(z, 0, 0) = \frac{\beta_1[1 - \tilde{R}_1(\lambda - \lambda z)]}{\lambda - \lambda z} \tilde{P}_1(z, 0), \\ \tilde{\tilde{P}}_4(z, 0, 0) = \frac{\beta_2[1 - \tilde{R}_2(\lambda - \lambda z)]}{\lambda - \lambda z} \tilde{P}_2(z, 0). \end{cases} \quad (8)$$

The pgfs of the orbit occupancy size, $P(z)$, and the system size, $R(z)$, can be straightforwardly written in terms of the partial generating functions of (8) as given below:

$$\begin{aligned} P(z) &= P_0(z) + \tilde{P}_1(z, 0) + \tilde{P}_2(z, 0) + \tilde{\tilde{P}}_3(z, 0, 0) + \tilde{\tilde{P}}_4(z, 0, 0), \\ R(z) &= P_0(z) + z[\tilde{P}_1(z, 0) + \tilde{P}_2(z, 0) + \tilde{\tilde{P}}_3(z, 0, 0) + \tilde{\tilde{P}}_4(z, 0, 0)]. \end{aligned} \quad (9)$$

From (8), the closed-form expressions for (9) can be easily obtained. By ignoring the non-zero probability of server failure, the following results are in complete agreement with [8]:

$$\begin{aligned} P(z) &= \frac{\lambda(1 - z) + \alpha[1 - \delta_2(z)]}{\lambda[\delta_1(z) - z]} P_0(z), \\ R(z) &= \frac{(\lambda - \lambda z)\delta_1(z) + \alpha[1 - \delta_2(z)]}{\lambda[\delta_1(z) - z]} P_0(z). \end{aligned} \quad (10)$$

4. Performance Metrics

The four performance measures are defined in this section.

4.1. Expected Number of Calls in the System

This measure accounts for the mean number of incoming calls retrying for service, either due to server failure or it being busy, as well as those being served by the server. This is readily obtained by differentiating the equations in (10) and evaluating them at $z = 1$. The first equation yields the first moment of the orbit size ($E[N]$) as follows:

$$\begin{aligned} E[N] = P'(1) &= \frac{\lambda^2[\beta_1\mu_{1,1}\gamma_{1,2} + (1 + \beta_1\gamma_{1,1})^2\mu_{1,2}]}{2[1 - \rho(1 + \beta_1\gamma_{1,1})]} \\ &+ \frac{\lambda\alpha[\beta_2\mu_{2,1}\gamma_{2,2} + (1 + \beta_2\gamma_{2,1})^2\mu_{2,2}]}{2[1 + \sigma(1 + \beta_2\gamma_{2,1})]} + \frac{\lambda[\rho(1 + \beta_1\gamma_{1,1}) + \sigma(1 + \beta_2\gamma_{2,1})]}{\nu[1 - \rho(1 + \beta_1\gamma_{1,1})]}, \end{aligned}$$

(11)

where $\rho = \lambda\mu_{1,1}$ and $\sigma = \alpha\mu_{2,1}$. Similarly, the second equation in (10) gives the following mean system size ($E[M]$):

$$E[M] = R'(1) = P'(1) + \frac{\rho(1 + \beta_1\gamma_{1,1}) + \sigma(1 + \beta_2\gamma_{2,1})}{1 + \sigma(1 + \beta_2\gamma_{2,1})}. \quad (12)$$

4.2. Server Availability

The probability that the server is operational at a given time instant t is defined as its point-wise availability, $A(t)$, and its steady state availability (i.e. $\lim_{t \rightarrow \infty} A(t) = P_a$) is given as follows:

$$\begin{aligned} P_a &= \lim_{z \rightarrow 1} \{P_0(z) + \tilde{P}_1(z, 0) + \tilde{P}_2(z, 0)\} \\ &= \frac{(1 + \sigma)[1 - \rho(1 + \beta_1\gamma_{1,1})] + \rho[1 + \sigma(1 + \beta_2\gamma_{2,1})]}{1 + \sigma(1 + \beta_2\gamma_{2,1})}. \end{aligned} \quad (13)$$

75 4.3. Server failure frequency

This corresponds to the probability that the server fails at time $t > 0$ given that it was operating at $t = 0$ [10]. The following closed-form results from (9):

$$P_f = \lim_{z \rightarrow 1} \{\beta_1 \tilde{P}_1(z, 0) + \beta_2 \tilde{P}_2(z, 0)\} = \rho\beta_1 + \sigma\beta_2 \frac{[1 - \rho(1 + \beta_1\gamma_{1,1})]}{[1 + \sigma(1 + \beta_2\gamma_{2,1})]}. \quad (14)$$

4.4. Expected Waiting Time in Orbit

Denoted by W , the steady-state delay experienced by a customer in orbit depends on the total idle time of the server not serving an incoming call (W_0), the total service time (including the server failure time) of the server providing service to an incoming call (W_1), and the total service time (including the server failure time) of the server busy with an outgoing call (W_2). The probability of an inbound call entering the orbit (P_w) is thus, given as follows:

$$\begin{aligned} P_w &= \lim_{z \rightarrow 1} \{\tilde{P}_1(z, 0) + \tilde{P}_2(z, 0) + \tilde{\tilde{P}}_3(z, 0, 0) + \tilde{\tilde{P}}_4(z, 0, 0)\} \\ &= \frac{\rho(1 + \beta_1\gamma_{1,1}) + \sigma(1 + \beta_2\gamma_{2,1})}{1 + \sigma(1 + \beta_2\gamma_{2,1})}. \end{aligned} \quad (15)$$

Using (15) and the first moments of the pgfs in (8), we obtain the mean waiting time in the orbit to be [5]:

$$E[W] = E[W_0] + E[W_1] + E[W_2] = E[N]/\lambda, \quad (16)$$

where $E[W_0] = P_w/\nu$, $E[W_1] = E[N]E[B_1] + \rho(1 + \beta_1\gamma_{1,1})E[R_1]$, and $E[W_2] = \sigma(1 + \beta_2\gamma_{2,1})(1 - \rho(1 + \beta_1\gamma_{1,1}))E[R_2]/(1 + \sigma(1 + \beta_2\gamma_{2,1})) + \sigma(1 + \beta_2\gamma_{2,1})E[W_0]$. Here, $E[B_i]$ and $E[R_i]$ represent the mean service time (including failure time) and the mean remaining service time (including failure time) while serving i -type calls, which are given as $\mu_{i,1}(1 + \beta_i\gamma_{i,1})$ and $E[B_i^2]/(2E[B_i])$, respectively.

80

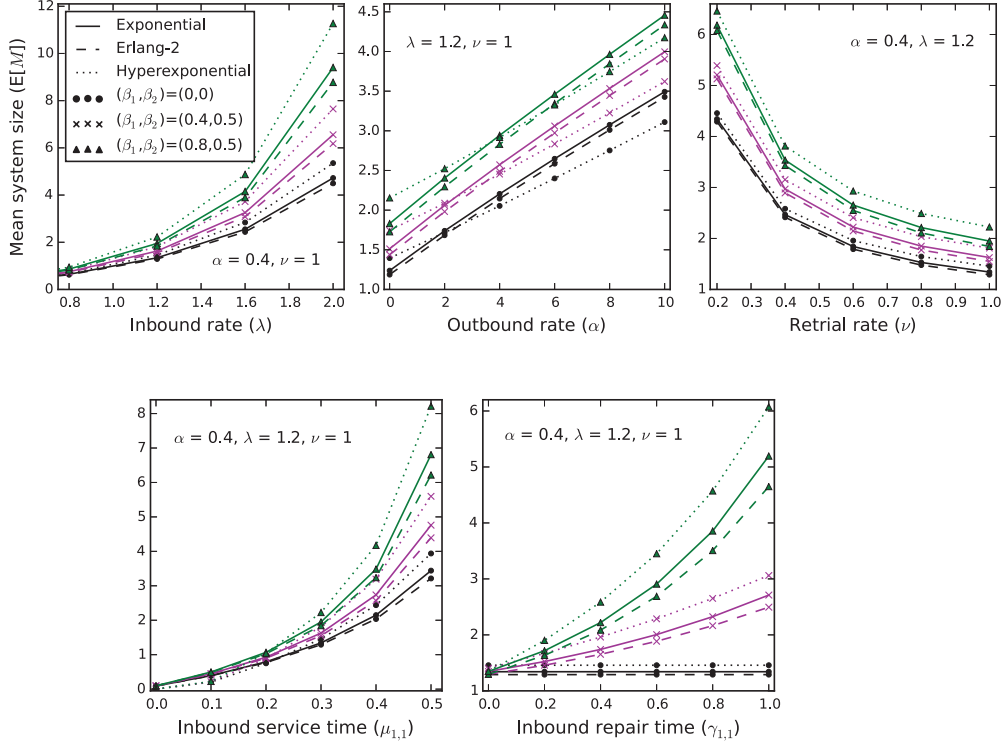


Figure 2: Mean system size ($E[N]$) versus incoming arrival rate (λ), outgoing rate (α), retrial rate (ν), inbound service time ($\mu_{1,1}$), and inbound repair time ($\gamma_{1,1}$).

5. Numerical Examples and Discussions

To illustrate the impact of system parameters on the performance primitives, we present numerical examples for service and repair times with three arbitrary distributions namely, exponential with density function $c_1 e^{-c_1 x}$, Erlangian of order two with density function $c_1^2 x e^{-c_1 x}$ and hyperexponential given as $a c_1 e^{-c_1 x} + (1-a) c_2 e^{-c_2 x}$, where $c_1, c_2 > 0$ and $0 \leq a \leq 1$. Throughout this section, we assume $\lambda = 1.2$, $\alpha = 0.4$, $\nu = 1$, $\mu_{2,1} = 0.1$, and $\gamma_{2,1} = 0.2$ to satisfy the ergodic condition of the system. We also consider an $M/G/1$ retrial queue without server failure [8], i.e. $(\beta_1, \beta_2) = (0, 0)$ as the baseline for comparison.

Figure 2 shows the variation in mean system size as functions of the inbound arrival rate (λ), outbound rate (α), retrial rate (ν), and inbound service ($\mu_{1,1}$) and repair ($\gamma_{1,1}$) times. Increase in the number of arriving calls reduces the chances of finding the server active and idle. Consequently, these calls enter the orbit to retry for service as shown in the figure. In comparison to the failure-free scenario, we note that the system size in our model increases with the failure rate β_1 as λ increases. A similar relationship can be observed between $E[M]$ and α as well. In regard to the retrial rate, $E[M]$ steeply decreases initially

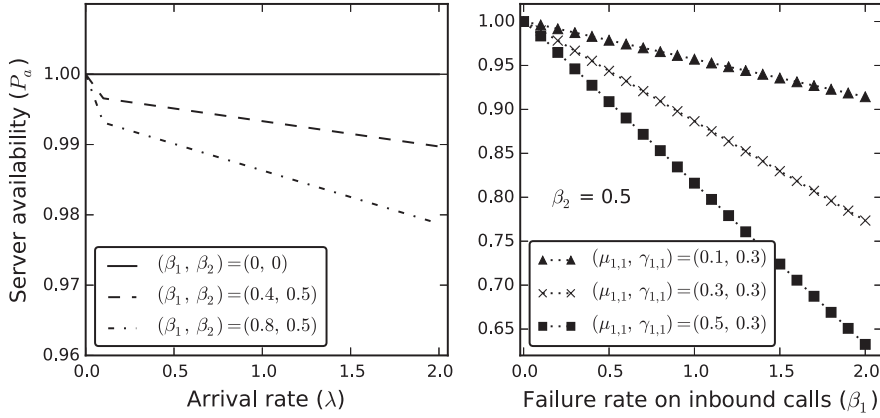


Figure 3: Server availability (P_a) versus inbound arrival rate (λ) and server failure rate while serving such calls (β_1).

and gradually stabilizes to some constant value. This justifies the fact that the calls residing in orbit have higher chances of being served and thus, leaving the orbit if they retry for service more frequently. For lower values of β_1 , the primary incoming calls are more probable to find the server available, resulting in a reduced orbit size. The influence of the service and repair times of primary calls is also evident in this figure. As the average time to serve incoming calls increases, $E[M]$ grows steeper. In other words, longer service times increases the number of incoming calls waiting in the orbit. Likewise, shorter the server repair time, the more active it would be thus, reducing the system size which is mainly dominated by the orbit length.

Figure 3 depicts the impact of λ and β_1 on server availability. Note that all three distributions exhibit the same results for different values of β_1 and β_2 . In absence of server failure, P_a is always 1. However, as β_1 increases, the availability of the server to incoming calls reduces with rise in λ . For instance, at $\lambda = 1.6$, P_a falls by slightly less than 1% as β_1 goes from 0.4 to 0.8. The figure also portrays the effect of β_1 under varying first moment values of service and repair times. Note that there is a steeper fall in P_a as the service time of incoming calls increases. Thus, the probability of finding the system available is higher when $\mu_{1,1} < \gamma_{1,1}$ and decreases with increase in $\mu_{1,1}$.

Similarly, Figure 4 describes the server failure frequency in terms of λ and β_1 . Apparent from (14), we see that P_f monotonically increases with the number of incoming calls in our model. Additionally, at $\lambda = 2$, as β_1 increases from 0.4 to 0.8, P_f rises drastically by over 74%. The profound contribution of (β_1, β_2) is also reflected in this figure. For various values of $(\mu_{1,1}, \gamma_{1,1})$, W_f increases constantly with β_1 and is higher when $\mu_{1,1}$ is greater than $\gamma_{1,1}$.

The mean orbit waiting time as functions of λ and β_1 are illustrated in Figure 5. With respect to the benchmark, we observe the impact of server failure while serving incoming calls on the gradual increase in $E[W]$. As β_1

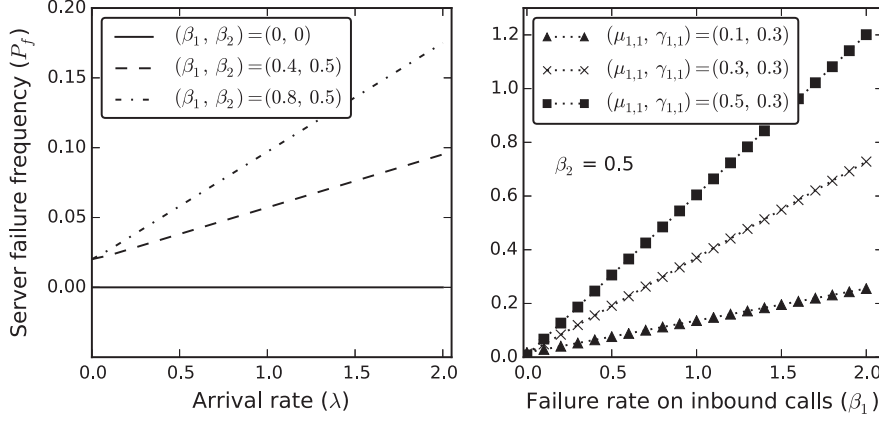


Figure 4: Server failure frequency (P_f) versus inbound arrival rate (λ) and server failure rate while serving such calls (β_1).

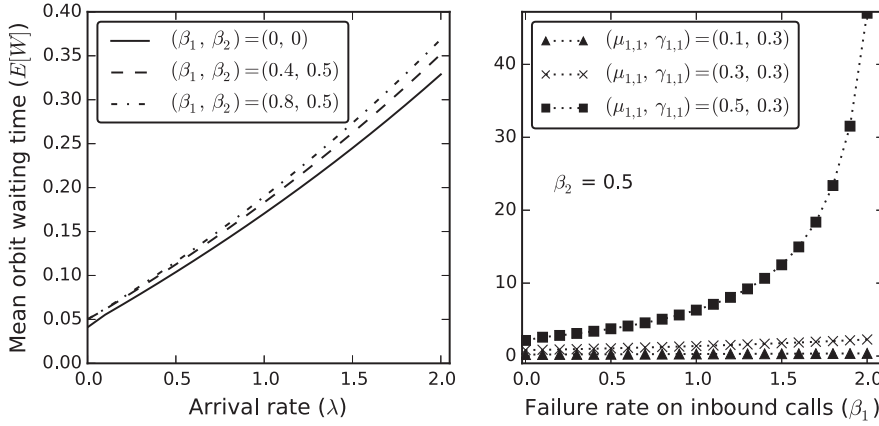


Figure 5: Expected waiting time in orbit ($E[W]$) versus inbound arrival rate (λ) and server failure rate while serving such calls (β_1).

increases to 2, the average orbit delay grows exponentially with increase in $\mu_{1,1}$.

6. Conclusion

In this paper, we modelled call blending in call centers as an $M/G/1$ retrial queue with the possibility of server failure and studied the system behavior in steady-state. In our analysis, we derived the system stability condition and closed-form expressions for the joint distribution of the server state and the expected number of customer calls in the system. Numerical results were provided for various performance measures to validate and compare our findings with that of a system with no breakdown.

Acknowledgements

This research was a part of the project entitled ‘Domestic Products Development of Marine Survey and Ocean Exploration Equipments’, and ‘Development of an Integrated Control System of Eel Farms based on Short-range Wireless Communication’, funded by the Ministry of Oceans and Fisheries, South Korea.

References

References

- [1] S. Bhulai, G. Koole, A queueing model for call blending in call centers, *IEEE Transactions on Automatic Control* 48 (8) (2003) 1434–1438. doi:10.1109/TAC.2003.815038.
- [2] Z. Aksin, M. Armony, V. Mehrotra, The modern call center: a multi-disciplinary perspective on operations management research, *Production and Operations Management* 16 (6) (2007) 665–688. doi:10.3401/poms.
- [3] N. P. Sherman, J. P. Kharoufeh, An M/M/1 retrial queue with unreliable server, *Operations Research Letters* 34 (6) (2006) 697–705. doi:http://dx.doi.org/10.1016/j.orl.2005.11.003.
- [4] J. Artalejo, A. Gomez-Corral, *Retrial Queueing Systems: A Computational Approach*, Springer-Verlag Berlin Heidelberg, 2008.
- [5] B. Choi, K. Choi, Y. Lee, M/G/1 retrial queueing systems with two types of calls and finite capacity, *Queueing Systems* 19 (1-2) (1995) 215–229. doi:10.1007/BF01148947.
- [6] J. Artalejo, J. Resing, Mean value analysis of single server retrial queues, *Asia-Pacific Journal of Operational Research* 27 (3) (2010) 335–345. doi:10.1142/S0217595910002739.
- [7] J. Artalejo, T. Phung-Duc, Markovian retrial queues with two way communication, *Journal of Industrial and Management Optimization* 8 (4) (2012) 781–806. doi:http://dx.doi.org/10.3934/jimo.2012.8.781.
- [8] J. Artalejo, T. Phung-Duc, Single server retrial queues with two way communication, *Applied Mathematical Modelling* 37 (4) (2013) 1811–1822. doi:http://dx.doi.org/10.1016/j.apm.2012.04.022.
- [9] L. I. Sennott, P. A. Humblet, R. L. Tweedie, Mean drifts and the non-ergodicity of markov chains, *Operations Research* 31 (4) (1983) 783–789. doi:http://dx.doi.org/10.1287/opre.31.4.783.
- [10] M. S. Kumar, R. Arumuganathan, An $M^X/G/1$ retrial queue with two-phase service subject to active server breakdown and two types of repair, *International Journal of Operational Research* 8 (3) (2010) 261–291. doi:10.1504/IJOR.2010.033540.